

## METHODS AND APPARATUS FOR PERFORMING SPEECH RECOGNITION AND USING SPEECH RECOGNITION RESULTS

The present invention is directed to speech  
5 recognition techniques and, more particularly, to methods  
and apparatus for generating speech recognition models,  
distributing speech recognition models and performing  
speech recognition operations, e.g., voice dialing and  
word processing operations, using speech recognition  
10 models.

Speech recognition, which includes both speaker  
15 independent speech recognition and speaker dependent  
speech recognition, is used for a wide variety of  
applications.

Speech recognition normally involves the use of  
20 speech recognition models or templates that have been  
trained using speech samples provided by one or more  
individuals. Commonly used speech recognition models  
include Hidden Markov Models (HMMS). An example of a  
common template is a dynamic time warping (DTW) template.  
25 In the context of the present application "speech

5           As part of a speech recognition operation,  
speech input is normally digitized and then processed.  
The processing normally involves extracting feature  
information, e.g., energy and/timing information, from  
the digitized signal. The extracted feature information  
10 normally takes the form of one or more feature vectors.  
The extracted feature vectors are then compared to one or  
more speech recognition models in an attempt to recognize  
words, phrases or sounds.

Before speech recognition operations can be performed, one or more speech recognition models need to be trained. Speech recognition models can be either speaker dependent or speaker independent. Speaker dependent (SD) speech recognition models are normally trained using speech from a single individual and are designed so that they should accurately recognize the speech of the individual who provided the training speech but not necessarily other individuals. Speaker independent (SI) speech recognition models are normally

generated from speech provided from numerous individuals or from text. The generated speaker independent speech recognition models often represent composite models which take into consideration variations between different speakers, e.g., due to differing pronunciations of the same word. Speaker independent speech recognition models are designed to accurately identify speech from a wide range of individuals including individuals who did not provide speech samples for training purposes.

10

In general, model training involves one or more individuals speaking a word or phrase, converting the speech into digital signal data, and then processing the digital signal data to generate a speech recognition model. Model training frequently involves an iterative process of computing a speech recognition model, scoring the model, and then using the results of the scoring operation to further improve and retrain the speech recognition model.

20

Speech recognition model training processes can be very computationally complex. This is true particularly in the case of SI models where audio data from numerous speakers is normally processed to generate each model. For this reason, speech recognition models are often generated using a relatively powerful computer systems.

25

05726971 "113000

0 In addition to limits in storage capacity,  
portable devices are often equipped with limited  
processing power. Speech recognition, like the model  
training process, can be a relatively computationally  
complex process and can therefore be time consuming given  
5 limited processing resources. Since most users of a  
speech processing system expect a prompt response from  
the system, to satisfy user demands speech processing  
often needs to be performed in real or near real time.  
As the number of potential words which may be recognized  
0 increases, so does the amount of processing required to  
perform a speech recognition operation. Thus, devices  
with limited processing power which may be able to  
perform a speech recognition operation involving  
recognizing, e.g., 20 possible names in near real time,  
5 may not be fast enough to perform a recognition operation  
in near real time where the number of names is increased  
to 100 possible names.

In the case of voice dialing and other applications where the recognition results need to be generated in near real time, e.g., with relatively little delay, the limited processing power of portable devices often limits the size of the vocabulary which can be considered as possible recognition outcomes.

In addition to the above implementation problems, implementers of speech recognition systems are often confronted with logistical problems associated with collecting speech samples to be used for model training purposes. This is particularly a problem in the case of speaker independent speech recognition models where the robustness of the models are often a function of the number of speech samples used for training and the differences between the individuals providing the samples. In applications where speech recognition models are to be used over a wide geographical region, it is particularly desirable that speech samples be collected from the various geographic regions where the models will ultimately be used. In this manner, regional speech differences can be taken into account during model training.

Another problem confronting implementers of speech recognition systems is that older speech recognition models may include different feature information than current speech recognition models. When updating a system to use newer speech recognition models,

5

10

20

where different types of speech recognition models are used by different speech recognizers is also desirable. Enhanced methods and apparatus for updating speech recognition models are also desirable.

5

#### SUMMARY OF THE INVENTION

09726971.1.3000  
The present invention is directed to methods  
10 and apparatus for generating, distributing, and using  
speech recognition models. In accordance with the  
present invention, a shared, e.g., centralized, speech  
processing facility is used to support speech recognition  
for a wide variety of devices, e.g., notebook computers,  
15 business computer systems personal data assistants, etc.  
The centralized speech processing facility of the present  
invention may be located at a physically remote site,  
e.g., in a different room, building, or even country,  
than the devices to which it provides speech processing  
20 and/or speech recognition services. The shared speech  
processing facility may be coupled to numerous devices  
via the Internet and/or one or more other communications  
channels such as telephone lines, a local area network  
(LAN), etc.

25

In various embodiments, the Internet is used as the communications channel via which model training data is collected and/or speech recognition input is received by the shared speech processing facility of the present

5 recognized words or phrases included in the processed speech. The speech recognition models may be returned as E-mail message attachments while the recognized words may be returned as text in the body of an E-mail message or in a text file attachment to an E-mail message.

Thus, via the Internet, devices with audio capture capability and Internet access can record and transmit to the centralized speech processing facility of the present invention digitized speech, e.g., as speech files. The speech processing facility then performs a model training operation or speech recognition operation using the received speech. A speech recognition model or data message including the recognized words, phases or other information is then returned depending on whether a model training or recognition operation was performed, to the device which supplied the speech.

Thus, the speech processing facility of the present invention can be used to provide speech recognition capabilities and/or to augment a device's speech processing capability by performing speech recognition model training operations and/or additional speech recognition operations which can be used to supplement local speech recognition attempts.



15

20

25

5

As discussed above, in various embodiments, the remote speech processing facility of the present invention is used to perform speech recognition operations and then return the recognition results or take other actions based on the recognition results. For example, in one embodiment business computer systems capture speech from, e.g., customers, and then transmit the speech or extracted speech information to the shared speech processing facility via the Internet. The remote speech processing facility performs speech recognition operations on the received speech and/or received extracted speech information. The results of the recognition operation, e.g., recognized words in the form of, e.g., text, are then returned to the business computer system which supplied the processed speech or speech information. The business system can then use the information returned by the speech processing facility, e.g., recognized text, to fill in forms or perform other services such as automatically respond to verbal customer inquiries. Thus, the remote speech processing method of the present invention can be used to supply speech processing capabilities to customers, e.g., businesses, who can't, or do not want to, support local speech processing operations.

In addition to providing speech recognition capabilities to systems which can't perform speech recognition locally, the speech processing facility of the present invention is used in various embodiments to augment the speech recognition capabilities of various devices such as notebook computers and personal data assistants. In such embodiments the remote speech processing facility may be used to perform speech recognition when the local device is unable to obtain a satisfactory recognition result, e.g., because of a limited vocabulary or limited processing capability.

In one particular exemplary embodiment, a notebook computer attempts to perform a voice dialing operation on received speech using locally stored speech recognition models prior to contracting the speech processing facility of the present invention. If the local speech recognition operation fails to result in the recognition of a name, the received speech or extracted feature information is transmitted to the remote speech processing facility. If the local notebook computer can't perform a dialing operation the notebook computer also transmits to the remote speech processing facility a telephone number where the user of the notebook computer can be contacted by telephone. The remote speech processing facility performs a speech recognition operation using the received speech and/or extracted feature information. If the speech recognition operation

5 requesting that the voice dialing speech recognition  
operation be performed unless a contact telephone number  
was provided with the speech and/or extracted feature  
information. In such a case, the speech processing  
facility uses telephone circuitry to initiate one  
10 telephone call to the telephone number retrieved from  
memory and another telephone call to the received contact  
telephone number. When the two calls are answered, they  
are bridged thereby completing the voice dialing  
operation.

In addition to generating new speech recognition models to be used in speech processing operations and providing speech recognition services, the centralized speech processing facility of the present invention can be used for modernizing existing speech recognition system but upgrading speech recognition models and the speech recognition engine used therewith. In one particular embodiment, speech recognition models or templates are received via the Internet from a system to be updated along with speech corresponding to the modeled words. The received models or templates and/or speech are used to generate updated models which include different speech characteristic information or have a

different model format than the existing speech recognition models. The updated models are returned to the speech recognition systems along with, in some cases, new speech recognition engine software.

5

In one particular embodiment, speech recognition templates used by voice dialing systems are updated and replaced with HMMs generated by the central processing system of the present invention.

10

At the time the templates are replaced, the speech recognition engine software is also replaced with a new speech recognition engine which uses HMMs for recognition purposes.

15

Various additional features and advantages of the present invention will be apparent from the detailed description which follows.

20 BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates a communication system implemented in accordance with an exemplary embodiment of the present invention.

25

Fig. 2 illustrates the communications system of Fig. 1 in greater detail.

09726971 "113000

Fig. 4 illustrates memory which may be used as the memory of a computer in the system illustrated in Fig. 1.

Fig. 6 illustrates a voice dialing IP device which may be used in the system illustrated in Fig. 1.

Fig. 8 illustrates an exemplary voice dialing routine of the present invention.

Fig. 10 illustrates a remote voice dialing routine implemented in accordance with the present invention.

Fig. 11 illustrates a call establishment routine of the present invention.

5                    Fig. 13 illustrates a speech processing  
facility implemented in accordance with one embodiment of  
the present invention.

### DETAILED DESCRIPTION

Fig. 1 illustrates a communications system 100 implemented in accordance with the present invention. As illustrated, the system 100 includes a business premises 10 and customer premises 12, 14, 16. Each one of the premises 10, 12, 14, 16 represents a customer or business site. While only one business premise 10 is shown, it is to be understood that any number of business and customer premises may be included in the system 100. The various

5 subscriber lines (DSL), cable modems, cable lines, high  
speed data links, e.g., T1 links, dial-up lines, wireless  
connections or a wide range of other communications  
channels. The premises 10, 12, 14, 16, 18 may be  
connected to the speech processing facility via a LAN or  
10 other communications channel instead of, or in addition  
to, the Internet.

high speed services, which may be on during all hours of business service, residential customers are now also moving to relatively high speed Internet connections which are "always on". As part of such services, a link to the Internet is maintained while the computer user has his/her computer on avoiding delays associated with establishing an Internet connection when data needs to be sent or received over the Internet. Examples of such Internet connections include cable modem services and DSL services. Such services frequently support sufficient bandwidth for the transmission of audio signals. As the speed of Internet connections increases, the number of Internet service subscribers capable of transmitting audio signals in real or near real time will continue to increase.



5  
10  
15

20

25

5

10

25

VD IP 70 is coupled to the Internet 30 via a network interface 72 and to the first switch 74 via a voice and signaling connection. VD IP 70 includes circuitry for performing voice dialing operations. Voice dialing operations include speech recognition operations and the placing of a call in response to the outcome of a speech recognition operation. Voice dialing IP 70 may include, for each voice dialing service subscriber supported by the VD IP 70, a voice dialing directory which includes speech recognition models of names of people who may be called, with associated telephone numbers to be dialed when the name is recognized.

Fig. 3 illustrates the computer system 50 which may be used at one or more customer premises, in greater detail. The computer 50 may be, e.g., a personal computer (PC), notebook computer, or personal data assistant (PDA). As illustrated the computer 50 includes memory 302, a processor 304, display device 314, input

While not illustrated in Fig. 3, in the case where

5 wireless Internet access is supported, modem 320 may be coupled to antenna 52 shown in Fig. 2.

Processor 304, under direction of routines stored in memory 302, controls the operation of the computer 50. Information and data may be displayed to a user of the device 50 via display 314 while data may be manually entered into the computer via input device, e.g., keyboard 316. The NIC 318 can be used to couple the computer 50 to a local area network (LAN) or other computer network. Modem 320 may be, e.g., a DSL modem, cable modem or other type of modem which can be used to connect the computer system to the Internet 30. Thus, via modem 320 the computer 50 can receive data from, and transmit data to, other devices coupled to the Internet 30.

To provide the computer system 50 with the ability to perform various telephone functions such as dial a telephone number and host telephone calls, the  
25 computer system 50 includes telephony circuit 308. An audio input device, e.g., microphone 310, provides audio input to the telephone circuit as well as audio signal processing circuitry 322. An audio output device, e.g., speaker 306, allows a user of the system to hear audio

5

10

20

25

5

15

20

5 Alternatively, an audio version of the recognized name  
may be generated from the text version 502 of the  
recognized name for confirmation message purpose.

In addition to the name and telephone number information included in the voice dialing customer record 520, the record also includes information 520, e.g., a world wide web Internet address, identifying a remote speech processing facility to be used in the event that a match is not identified between the models in the record and spoken speech being processed for voice dialing purposes or in the event that speech recognition models are to be updated or generated. The memory also includes a contact telephone number 522 where the user can be reached when the computer system's telephone connection is not enabled.

When the voice dialing customer record 520 includes speaker dependent speech recognition models, it may be used as the SD voice dialing customer record 422 shown in Fig. 4. When the voice dialing customer record 520 includes speaker independent speech recognition models, it may be used as the SD voice dialing customer record 424.

As illustrated the speech processing system 18 includes memory 1302, a processor 1304, display device 1314, input device 1316, telephony/call initiation circuit 1308, network interface card (NIC) 1318, modem 1320 and audio signal processing circuitry 1322 which are coupled together via bus 1313. Processor 1304, under direction of routines stored in memory 1302, controls the operation of the system 18. Information and data may be displayed to a system administrator via display 1314 while data may be manually entered into the system 18 via input device 1316. The NIC 1318 can be used to couple the system to a local area network (LAN) or other computer network. Modem 1320 may be, e.g., a DSL modem, cable modem or other type of modem which can be used to connect the computer system to the Internet 30. Thus, via modem 1320 the system 18 can receive data from, and

As illustrated the speech processing system 18 includes memory 1302, a processor 1304, display device 1314, input device 1316, telephony/call initiation circuit 1308, network interface card (NIC) 1318, modem 1320 and audio signal processing circuitry 1322 which are coupled together via bus 1313. Processor 1304, under direction of routines stored in memory 1302, controls the operation of the system 18. Information and data may be displayed to a system administrator via display 1314 while data may be manually entered into the system 18 via input device 1316. The NIC 1318 can be used to couple the system to a local area network (LAN) or other computer network. Modem 1320 may be, e.g., a DSL modem, cable modem or other type of modem which can be used to connect the computer system to the Internet 30. Thus, via modem 1320 the system 18 can receive data from, and



transmit data to, other devices coupled to the Internet  
30.

To provide the system 18 with the ability to perform various telephone functions such as dial a telephone number and bridge telephone calls, the system 18 includes telephony/call initiation circuit 1308.

In order to support speech recognition model training, and speech recognition operations audio signal processing circuitry 1322 is provided. Processing circuitry 1322 includes a feature extractor circuit 1324, a speech recognition circuit 1328, and a model training circuit 1330 which are all coupled to bus 1313. Thus, the components of the audio signal processing circuitry 1322 can receive audio signals and extracted speech feature information via bus 1313. Extracted feature information, received speech, and generated speech recognition models can be stored in memory 1302. Memory 1302 is also used to store various routines and data used by the various components of the system 18.

The contents of the memory 1302 may include voice dialing data including voice dialing customer records for multiple customers. The memory 1302 also includes various speech recognition, call initiation and model training routines. In addition, the memory 1302 includes a training database 1209 which is a collection of speech samples used for training speech recognition

5 includes, e.g., system identification and contact  
information such as an E-mail address, the type of speech  
recognition models used by the individual systems, the  
words in each systems' speech recognition vocabulary, and  
information on when to update the each systems speech  
10 recognition models.

While speech processing facility 18 can support a wider range of speech processing operations including voice dialing, specific telephone switch peripheral devices such as VD IP 70 may be dedicated to supporting voice dialing operations. An exemplary voice dialing IP 70 which may be used as the voice dialing IP of Fig. 2 is shown in detail in Fig. 6. The VD IP 70 can support voice dialing operations in response to speech received via a conventional telephone connection or via the Internet 30. Thus, the computer system 50 can use the VD IP 70 to perform a voice dialing operation. This can be done by E-mailing the VD-IP 70 a voice dialing request message including speech in an attached file.

15

Speech recognizer 602 is coupled to a switch, e.g., SSP and receives voice signals therefrom. The speech recognizer circuit 602 uses speech recognition models stored in the memory 612 and the speech recognition routine 613 to perform a speech recognition operation on audio signals received from a telephone switch or from the Internet via network interface 610. Speech recognition models used by the speech recognizer 602 may be speaker independent and/or speaker dependent models. The speech recognition models are retrieved from the personal dialer and corporate records 618, 620 based on a customer identifier which identifies the particular customer whose speech is to be processed.

5 recognized. If a name is recognized, and speech was  
received via the Internet, the telephone number  
corresponding to the recognized name is returned via the  
Internet to the device which provided the speech.  
However, if a contact telephone number was received via  
the Internet with the speech to be processed, the voice  
0 dialing routine 614 calls the call setup routine 615  
which is responsible for imitating a call to the  
telephone number corresponding to the recognized name.

15  
20  
25

Instead of using VD IP 70, computer system 50 can use the speech processing facility 18 to support a voice dialing operation. Voice dialing will now be described from the perspective of computer system 50 as

it interacts with speech processing facility 18. Fig. 8 illustrates an exemplary voice dialing routine 416 which may be executed by the computer system 50.

5           The voice dialing routine 800 begins in start  
step 802 when it is executed, e.g., by the processor 305  
of computer system 50. From step 802, operation proceeds  
to step 804 wherein the routine monitors for speech  
input. If in step 806, it is determined that speech was  
10 received in step 804, operation proceeds to step 808.  
Otherwise, operation returns to monitoring step 804.

In step 808 a determination is made as to whether or not local speech feature extraction is supported. If it is not, operation proceeds directly to step 818. However, if local feature extraction is supported, e.g., feature extractor 324 is present, operation proceeds to step 810 wherein a feature extraction operation is performed on the received speech. Next in step 814 a determination is made as to whether or not local speech recognition capability is available, e.g., a determination is made whether or not the system includes speech recognition circuit 328. If in step 328 it is determined that local speech recognition is not available, operation proceeds directly to step 818. However, if local speech recognition capability is available, operation proceeds to step 812 wherein a local voice dialing sub-routine, e.g., the subroutine 900 illustrated in Fig. 9 is called.

10 In step 902, the subroutine is provided with the  
extracted feature information 903 produced, e.g., in step  
810, from the speech which is to be processed for voice  
dialing purposes. Operation then proceeds to step 904  
wherein a speech recognition operation is performed using  
15 the received extracted speech feature information and one  
or more locally stored speech recognition models, e.g.,  
speech recognition models obtained from the SD voice  
dialing customer record 422 or SI voice dialing customer  
record 424 stored in memory 302.

In step 906 a determination is made as to whether or not a name was recognized as a result of the voice dialing operation. If a name was not recognized operation proceeds to return step 908 wherein operation returns to step 812 of the voice dialing routine 800 with an indicator that the local voice dialing operation was unsuccessful.

5 connection exists. If the computer system 50 is connected to a telephone line, operation will proceed to step 914. In step 914, the computer system 50 is made to dial the telephone number associated, e.g., in one of the voice dialing records 422, 424, with the recognized name.

10 Then, in step 916, the computer system 50 detects completion of the call initiated in step 914 before proceeding to step 918.

From step 912 operation proceeds to return step 918. In return step 918 operation is returned to step

5

10

20

25



5

10

As will be discussed below, in response to the transmitted information, the speech processing facility 18 executes a voice dialing routine. Upon detecting the name of a party having an associated telephone number, the executed routine returns, e.g., in an E-mail message, the telephone number associated with the recognized name via the Internet assuming a contact telephone number was not provided to the facility 18. The telephone number can then be used by the computer system 50 to place a call to the party whose name was spoken. In the case where the computer system provides a contact telephone number to the speech processing system 18, the system 18 realizes that the computer 50 cannot place the call. In such a case, the remote speech processing facility 18 returns a signal indicating that the named party is being called assuming a name was recognized or that the system was unable to identify a party to be called in the event a name was not recognized.

5 is made as to whether or not the received response includes a telephone number to be dialed. If the response does not include a telephone number to be dialed, operation proceeds to step 829 where the system user is provided a message indicating the results of the remote voice dialing operation. That is, the system user is notified if the named party is being called or that the system was unable to identify a party to be called. The message to be provided is indicated by the response received from the speech processing facility 18.

5 Operation proceeds from notification step 829 via GOTO step 834 to monitoring step 804.

20 proceed from step 826 to step 830 wherein the computer system 50 dials the received telephone number. After call completion is detected in step 832, operation proceeds to step 804 via GOTO step 834. In this manner, the voice dialing routine returns to a state of

25 monitoring for speech input, e.g., input associated with an attempt to place another telephone call.

Voice dialing from the perspective of the speech processing facility will now be described with

Next, in step 1006, voice dialing information is retrieved from memory. The retrieved information may include, e.g., a voice dialing record including speech recognition models and corresponding telephone numbers to be used in providing voice dialing services for the identified user. The voice dialing record may be a customer specific record, e.g., part of a personal voice dialing record corresponding to the received user ID, or a common voice dialing record such as a corporate voice dialing directory shared by many individuals including the user identified by the received user ID.

10

15

20

*[Handwritten signature]*

number is transmitted to a call initiation device. The user's ID information may also be transmitted to the call initiation device. The call initiation device may be, e.g., conference calling IP 78 or circuitry interval to the speech processing system 18.

When the call initiation device is an external device such as conference calling IP 78, the telephone number to be dialed, the contact telephone number, and the user ID information is transmitted to the call initiation device over any one of a plurality of communication channels including the Internet, a LAN, and conventional telephone lines. In response to receiving the transmitted information the call initiation device executes a call establishment routine, e.g., the routine 1100 illustrated in Fig. 11, will initiate a call to both the telephone number to be dialed and the contact telephone number and then bridge the calls when they are answered. From step 1018 of Fig. 10, operation proceeds to step 1028.

In step 1016, of Fig. 10, if it is determined that a telephone contact number was not received, e.g., because the device which transmitted the voice dialing information is capable of initiating a call, operation proceeds to step 1020 wherein the telephone number to be dialed is transmitted (returned) to the remote computer system 50 in response to the received voice dialing

information, e.g., received speech and user ID information. Then operation proceeds to step 1028.

Referring once again to step 1014 if it is  
5 determined in this step that a name was not recognized by  
the speech recognition operation then processing proceeds  
to step 1022 instead of step 1016. In step 1022 a  
determination is made as to whether there is an  
additional remote speech processing system associated  
10 with the identified user, e.g., another system such as VD  
IP 70 which can be used support a voice dialing  
operation. This determination may be made by checking  
information about the user stored in memory.

15           If the answer to the inquiry made in step 1022  
is no, operation proceeds to notification step 1023 prior  
to proceeding to STOP step 1028. In step 1023 a message  
is sent back to the system 50 indicating to the system  
that the voice dialing attempt failed due to a failure to  
20 recognize a name.

If in step 1022 it is determined that there is an additional remote speech processing system associated with the identified user, operation will proceed from step 1022 to step 1024. In step 1024 the user ID information is transmitted to the additional remote speech processing facility associated with the identified user. Then, in step 1026, the previously received speech information and/or feature information is transmitted to

5

10

15

25

In step 1106 the conference calling IP initiates a call using the telephone number to be dialed while in step

15

In addition to supporting voice dialing operations, the speech processing 18 is capable of receiving speech signals, e.g., in compressed or uncompressed digital form, generating speech recognition  
20 models from the received speech, and then distributing the generated models to one or more devices, e.g., voice dialing IPs, business sites which perform speech recognition, and individual computer systems 50. In accordance with one feature of the present invention  
25 speech to be used in speech recognition model training operations, and the models generated there from, are transmitted over the Internet. Alternatively, other communications channels such as conventional telephone lines may be used for this purpose.



5

10

20

10

15

20

As will be discussed below, the speech processing facility 18 processes the transmitted speech

5 for storage and/or use in speech recognition operations.

10 models from the speech processing facility 18. The received speech recognition models will include the model or models generated from the speech extracted feature information and/or other information transmitted to the speech processing facility in step 714.

20 corresponding to the same words, names or sounds.

25 voice dialing and word processor applications. After storage of the received models, the new model training routine 700 then stops in step 720 until being executed again to train an additional model.

In step 1204 the system monitors for a model generation and/or model updating service request, e.g., a signal from a device such as the computer system 50 or computerized business system 58 indicating that a speech recognition model needs to be generated or updated. The request may take the form of an E-mail message with an attachment including information, speech and/or other speech data. When a request for such a service is received, e.g. via the Internet 30, operation proceeds to step 1206 wherein the information and data used to provide the requested service is received by the processor 1304, e.g., by extracting the attachment from the E-mail request message. The received information depends on the service to be performed.

Block 1206a illustrates exemplary data that is received with a request to generate a new speech recognition model. The data 1206a includes a User ID, speech or feature information, text information providing a text representation of the word or phrase to be modeled, and optional speech recognition model type information. The User Id may be a telephone number, E-mail address or some other type of unique identifier. Assuming model type information is not provided a default model type will be used.

Block 1206b illustrates exemplary data that is received with a request to update an existing speech recognition model. The data 1206b includes a User ID, an existing speech recognition model to be updated, existing model type information, speech or feature information, text information providing a text representation of the word or phrase to be modeled, and optional updated speech recognition model type information. If the optional updated speech recognition model type information is not provided, it is assumed that the updated model is to be of the same type as the received existing model.

Operation proceeds from step 1206 to step 1208. In step 1208, the training database 1209 maintained in the speech processing facility 18 is augmented with the speech received in step 1206. Thus, over time, the size and robustness of the speech training database 1211 will improve from the input received from various sources

5

10

20

the case of a speaker dependent speech recognition model

type, the generated model will be a speaker dependent speech recognition model. In the case of speaker independent speech recognition model the generated model will be a speaker independent model. Speaker independent models are normally trained using the received speech and speech included in the training database 1209 as training data. Speaker dependent models are normally generated using the received speech as the training data. In addition to indicating whether a generated model is to be speaker independent or speaker dependent the received model type information can indicate particular features or information which are to be used in the model, e.g., energy and delta energy coefficient information. In the case of models which are being updated, the updated model type information can specify a different model type than the existing model type information.

In one particular application, a dynamic time warping (DTW) template is received and processed along with speech to generate a speaker dependent Hidden Markov model as an updated model. In such an embodiment the received existing model type information would be e.g., "DTW template" and the updated model type information would be "SD HMM" indicating a speaker dependent HMM. In this particular application, the template to HMM model conversion and training techniques discussed in U.S. Patent No. 6,014,624 which is hereby expressly incorporated by reference may be used in the model generation step 1210.

From step 1212, operation proceeds to step 1214 wherein the generated speech recognition model is transmitted to the device from which the model generation or updating request was received. Operation then proceeds to step 1204 wherein the processor monitors for additional input, e.g., requests to generate or update additional speech recognition models.

25           The processing path which begins with step 1224  
executes in parallel with the processing path which  
beings with step 1204. In step 1224 a system clock is  
maintained. Operation proceeds from step 1224 to step  
1226 wherein a determination is made as to whether or not



As an alternative to broadcasting updated  
20 speech recognition models on a periodic basis, systems  
which use speech recognition models can periodically  
request, from the speech processing facility 18, speech  
recognition model updates via the Internet.

25           As discussed above, the speech processing facility 18 can be used to provide speech recognition services in addition to voice dialing and speech recognition model training services. Speech recognition service can be provided to devices, e.g., computer system

20

Fig. 14 illustrates a speech recognition routine that is implemented by the speech processing facility 18 to service speech processing requests received from various devices coupled to the Internet 30.

25 As illustrated, the routine 1400 begins in step 1402, wherein the routine 1400 is retrieved from memory 1302 and executed by the speech processing facility's processor 1304.

5

15

25

5

10

15